International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 03, Issue 09, September 2014 ISSN 2319 - 4847

Personalized Search Engine for Mobiles

Ankita Bung[1], Mateti.Jhansi[2]

[1]PG Scholar, Dept of CSE, Marri Laxman Reddy Institute of Technology, Dt: R.R District, AP, India

[2]Assistant Professor, Dept of CSE, Marri Laxman Reddy Institute of Technology, Dt: R.R District, AP, India

ABSTRACT

Searching web contents using mobiles are increasing day by day. The personalized search engine for mobile is a new client-server model for a mobile search engine for providing relevant information based on user profile. The PSEM done a search using both content based and location based information of the user. The location data captured using GPS. User preferences are maintained using ontology based criteria provided by PSEM server and this is used to train and re rank the new search results. The client, through data is also maintained at client side in the user profile for preserving the privacy. Extracting the concepts, training and re-ranking done on server side. The proposed system takes the re-ranked links to eliminate the duplicate links, so that we can improve the web search results. A GAHWM-WWBADS algorithm is used to remove the duplicated links. This task is also done at PSME server. WWBADS is used to remove duplicate links and GAHWM is used for ordering the links.

Keywords: PSEM, mobile search engine, concept and location search, ontology, webpage duplication.

1. INTRODUCTION

Mobile device usage has been increased for last 10 years. The searching the web contents is also using mobiles is also increased. As the basic limitations of mobile devices like hardware and software capabilities, it cannot huge data processing. It is very difficult for relevant information to the user query. The best solution to this problem is maintaining personalized web search. This can be done by taking user interests into the profile. These personalized interests can be captured and analyzed using users click through data. The Personalized web search engines are using this information to produce the more relevant data to the users search. Personalized search engine for mobile is a client-server model used to reduce the computational complexity at client mobile device and produces the most accurate and relevant results to the users search. Only client personalized information is maintained at client side. All the processing activities like concept extraction, training, re-ranking and elimination of duplication of search results done on server side. The personalized mobile search engine uses concept, principle for search results The concepts are further classified into two types content concept and location concept. Location information is captured from GPS of the mobile.

2. RELATED WORK

Click trough data is used to determine the user preferences on their search results. There are many existing personalized systems [1],[2] and [3] based on the click through data to find the user preferences. Search engines can be either content related or location related. According to Yokoji [4] we can search contents based on location. Location information was extracted from web pages. According to E. Agichtein, E. Brill, and S. Domains [5], we can improve the search results by taking the user feedback into account. G.Y. Chen et.. all [5] was developed a new geographic based search engine which is used to determine information based on the current location of the user. Research related to location based is very interesting and made researchers to think and apply new ideas on search engines. Kenneth Wai-Ting Leung, Dik Lun Lee, and Wang-Chien Lee [7] developed a new personalized mobile search engine to search the user queries. They use a client server model to reduce the complexity at user side and produce good results. It uses both content and location concepts to provide good results. In this paper, we are going to enhance the performance of PSME by eliminating the duplicate web pages in the results a server side. This can be done by using WWBADS algorithm. The reordering can be done using a genetic algorithm.

3. SYSTEM ARCHITECTURE

The architecture of the PSME was explained by the following diagram [7].

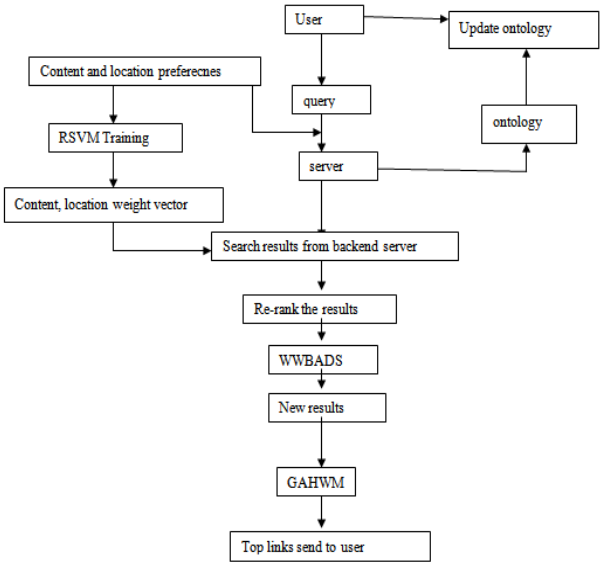

Figure 1: PSME architecture

PSME is a client server architecture. All searches related information and ontology data send by PSME server is stored in the client. All processing tasks like concept extraction, training, re-ranking and elimination of duplication is done on server side. After completion the server will send only required information to the user. The two basic tasks of PSME is to re-rank the web results using content weight vector and location weight vector taken from RSVM training and update the new ontology produced by the PSME server at client database. Here PSME four entropies called content-entropy, location entropy, content click entropy, location click entropy to improve the efficiency of the system.

4. PROPOSED SYSTEM

The new proposed system does not alter anything but it improves the search results by adding extra step at server side that is eliminating the duplicate content in the search results. We are using new algorithm GAHWM-WWBADS to eliminate the redundant pages. The following flow chart explains how PSME will work after adding this additional step at server side.

4.1 Re-ranking the search results

PSME server takes the backend search engine results and the concept and location weight vectors from RSVM training and re-rank according to the user preferences.

4.2 Updating the Ontology database

PSME server sends new ontology that describes the new relationships from the concepts used in the search results. This will be stored in the client space. Then click through data is also stored in the client database whenever he clicks a link.

4.3 Eliminating duplicated links

To eliminate the duplicate links, we are using the algorithm called WORD WEIGHTAGE BASED APPROACH FOR DOCUMENT SELECTION (WWBADS). After getting huge links, we reorder them by using GAHWS.

4.3.1 WWBADS algorithm

WORD WEIGHTAGE BASED APPROACH FOR DOCUMENT SELECTION (WWBADS) is used to eliminate the redundant links.
Input: Re-ranked links from RSVM algorithm
Output: Improved links by eliminating duplication.

Algorithm

Step1: Load the web documents.
Step2: Remove exact duplication documents.
Step3: Remove the nearly duplicated documents.
Step4: Reorder the links using GAHWM.
Step5: Display the resulted new documents.

4.3.2: GAHWM algorithm

Genetic Algorithm with ranking based objective function has been used to list the documents according to the user query. In our approach, we use the same fitness function to list the documents; however the provided list is free from the exact duplicates documents and nearly duplicate documents. The fitness function used in genetic algorithm is what determines whether a given solution is optimal or not. In genetic algorithms, solutions are represented as chromosomes. These chromosomes are modified in such ways that in each generation, the fitness value of these chromosomes gets closer the optimal solution. The chromosomes presented in this research contain a list of randomly chosen files. Chromosomes with high fitness value tend to be closer to the optimal solutions, thus making the fitness value of a chromosome determines whether the file is relevant or not. The program uses the fitness function presented from the research in web mining, Genetic Algorithm for HTML Web Content Mining (GAHWM).

$$F(C)=1/N \times \sum_{j=1}^{L} f(d) X \sum_{i=j}^{N} \frac{1}{L} \quad (1)$$

$$F(di) = \sum_{i=1}^{K} wj \quad (2)$$

$$W_i = \frac{Ki}{K} \times \frac{Fij}{Fj} \times \frac{1}{tj} \times h_{ij} \times \log\left(\frac{T}{Ti}\right) \times \log\left(\frac{N}{dfi}\right) \quad (3)$$

5. CONCLUSION

By using personalized mobile search we can get most user relevant results. The PSME will enhance the performance in the context of both results and device. Clearly, it will reduce the computational cost of client device. Ontology from sender will provides all the possibilities of content that can be used by the client. PSME server plays a major role to perform all the complex tasks. Our new proposed algorithm called GAHWM-WWBADS eliminates the all duplicate links on the final list. We can extend this algorithm to find the links during travelling that will change the user location dynamically.

REFERENCES

[1] E. Agichtein, E. Brill, S. Dumais, and R. Ragno, "Learning User Interaction Models for Predicting Web Search Result Preferences," Proc. Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 2006.

[2] T. Joachims, "Optimizing Search Engines Using ClickthroughData," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and DataMining, 2002.

[3] W. Ng, L. Deng, and D.L. Lee, "Mining User Preference Using SpyVoting for Search Engine Personalization," ACM Trans. InternetTechnology, vol. 7, no. 4, article 19, 2007.

[4] S. Yokoji, "Kokono Search: A Location Based Search Engine," Proc.Int'l Conf. World Wide Web (WWW), 2001.

[5] E. Agichtein, E. Brill, and S. Domains,"Improving web search ranking by incorporatinguser behaviour information," in Proc. of ACMSIGIRConference, 2006, pp.667-686**.**

[6] G.Y. Chen, T. Suel, and A. Markowetz, "Efficient query processing in geographic web search engines, "in Proc. of ACM SIGIR Conference,2006, pp.421-462.

[7] Kenneth Wai-Ting Leung, Dik Lun Lee, and Wang-Chien Lee "A personalize mobile search engine" in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 4, APRIL 2013.

[8] U. Li, Z. Li, W.C. Lee, and D. L. Lee, "A probabilistic topic-based ranking framework for location-sensitive domain information retrieval," in Proc. of ACM SIGIR Conference, 2009.

[9] M. Liu, W. S. Lee, P. S. Yu, and X. Li, "Partially supervised classification of text documents," in Proc.of ICML Conference, 2002.

[10] Li Zhiyi, Liyang Shijin, National Research on Deleting Duplicates Web Pages: Status and Summary, Library and Information Service,2011, 55(7): pp.118-121.

[11] Gaudence Uwamahoro, Zhang Zuping, Efficient Algorithm for Near Duplicate Documents Detection, International Journal Of Computer Science Issues, Vol 10, issue 2, March 2013.

[12] Metzler, D., Bernstein, Y., Croft, W.B., Moffat, A., Zobel,J., Similarity Measures for Tracking Information Flow, In: The 14th ACM Conference on Information and Knowledge Management (CIKM 2005), 2005, pp.517–524 [9] Zobel J, Alistair Moffat, Inverted files for Text Search Engines, ACM Computing Surveys, Vol . 38, No. 2, article 2006, pp. 1-55.

[13] Ajik Kumar Mahapatra, Sitanath Biswas, Inverted Index Techniques, International Journal of Computer Science Issues, Vol. 8, Issue 4, No. 1, 2011.